Joint User Grouping and Power Optimization for Secure mmWave-NOMA System

***Abstract*—Due to the proliferation of mobile devices, provisioning of massive connectivity has become a major challenge for future networks. The combination of millimeter wave (mmWave) with non-orthogonal multiple access (NOMA) provides a promising solution to massive connectivity. However, the security issue therein cannot be ignored due to the openness of wireless channels. To overcome the security challenge in mmWave-NOMA based networks, the nonorthogonal interference can be exploited to improve the security. In this paper, we propose a novel mmWave-NOMA framework where the users are classified as secure users (SUs) and common users (CUs), to satisfy their heterogeneous security service needs with the presence of randomly located eavesdroppers. According to their channel disparity, the NOMA users with stronger channel gains are deemed as SUs for better secrecy performance, while the remaining ones are served as CUs. To further enhance the security, hybrid precoding for SUs is designed to strengthen the desired signal and reduce interference. In addition, to reduce the complexity and satisfy the diverse demands, user grouping and power allocation are jointly optimized to maximize the sum rate of CUs subject to the SUs' requirements. To solve the intractable non-convex problem, we decompose it into two subproblems, i.e., user grouping and power optimization, and a hybrid SU-CU grouping algorithm and a successive convex approximation based algorithm are proposed to solve them, respectively. Finally, simulation results are provided to show the advantages of the proposed scheme.**

***Index Terms*—mmWave communications, non-orthogonal multiple access, power optimization, secrecy transmission, user grouping.**

I. Introduction

Power-domain non-orthogonal multiple access (NOMA) becomes a crucial multiple access (MA) technique to enhance the spectrum efficiency via power multiplexing for the fifth-generation mobile networks and beyond [2]. To achieve higher spectrum efficiency, the superposed signal of multiple NOMA users is transmitted by the base station (BS) over a single resource block, and successive interference cancellation (SIC) is utilized to detect the user messages according to different power levels at the destinations [3]. Following this nonorthogonal transmission and interference cancellation principle, plenty of research has been conducted to explore the application scenarios of NOMA systems, such as cooperative NOMA, millimeter wave (mmWave) NOMA, multi-input multi-output (MIMO) NOMA and cognitive radio inspired NOMA [4]. While different design aspects have been studied, most existing NOMA related research is focused on the communication aspect.

With the growing importance of communication security, we focus on the issue of the physical layer security (PLS) in NOMA networks. Similar to the conventional MA methods, NOMA is also susceptible to security threats due to the broadcast nature of wireless channels with potential malicious eavesdroppers. As a complementary technology to the upper-layer encryption algorithm, PLS aims at guaranteeing the confidential transmission of wireless networks via lightweight physical layer technologies such as signal processing and coding [5]. In [6], the technique based on PLS was used by Li *et al.* to improve the security of the multiuser orthogonal frequency division multiplexing network assisted by the full-duplex relay. The secure energy efficient was maximized by jointly considering subcarrier permutation, subcarrier pair allocation and power consumption under the assumption of imperfect eavesdropping channel state information (CSI). Yang and Xiong *et al.* investigated the secrecy rate maximization problem in [7] for the intelligent reflecting surface (IRS)-aided network with multiple eavesdroppers, where a novel deep learning based method was proposed to jointly optimize the active beamforming at the BS and the passive beamforming of the IRS considering the time-varying channel conditions. While for the NOMA network, there have been a number of works dedicated to enhancing its security through designing various anti-eavesdropping schemes from the view of PLS, such as artificial noise assisted schemes [8], [9], cooperative relaying/jamming strategies [10], [11], secure beamforming [12], [13] , secrecy analysis based on user pairing [14], *etc*. Specially, some of the existing research concludes that NOMA is beneficial to enhance the security due to its inherent power allocation (PA) and SIC. In [15], the enhancements of spectral efficiency and security were measured by Ding *et al.* for

Manuscript received April 26, 2021; revised August 16, 2021; accepted October 08, 2021. The work was supported bywas supported by the National Key R&D Program of China under Grant 2020YFB1807002, the National Natural Science Foundation of China (NSFC) under Grant 61871065, and the Fundamental Research Funds for the Central Universities. This paper will be presented in part at the Proceedings of IEEE GLOBECOM 2021 [1]. The associate editor coordinating the review of this paper and approving it for publication was X. Tao. *(Corresponding author: Nan Zhao.)*

Y. Cao, M. Jin and N. Zhao are with the Key Laboratory of Intelligent Control and Optimization for Industrial Equipment of Ministry of Education, Dalian University of Technology, Dalian, China (email: cy216@mail.dlut.edu.cn, mljin@dlut.edu.cn, zhaonan@dlut.edu.cn).

S. Wang is with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China (e-mail: swang@bit.edu.cn).

Y. Chen is with the School of Engineering, University of Warwick, Coventry CV4 7AL, U.K. (e-mail: Yunfei.Chen@warwick.ac.uk).

Z. Ding is with the School of Electrical and Electronic Engineering, The University of Manchester, Manchester, M13 9PL, U.K. (e-mail: zhiguo.ding@manchester.ac.uk).

X. Wang is with the Department of Electrical and Computer Engineering, University of Western Ontario, London, ON N6A 5B9, Canada (e-mail: xianbin.wang@uwo.ca).

the hybrid unicast-multicast streaming transmission in NOMA networks, where the secrecy rate of the unicasting stream was enhanced compared to orthogonal multiple access (OMA). In [16], the secrecy performance of NOMA with massive connections was investigated by Chen *et al.*, in which the harmful inter-user interference caused by NOMA was utilized to confuse the eavesdropping and improve the security. Xiao *et al.* proposed a hierarchical security based opportunistic NOMA schemes for massive MIMO to serve users with different security levels, and the throughput and secrecy capacity can be improved [17]. For the double unmanned aerial vehicle aided mobile edge computing network in [18], Xu and Zhang *et al.* aimed at minimizing the secure computing capacity maximization issues with the constrained UAVs' mobility and the communication and computation resources. Both cases of OMA and NOMA are considered, and it proves that the security of NOMA scheme is preferable to OMA scheme.

Although NOMA can enhance both the spectral efficiency and security of wireless communications, the hardware complexity caused by SIC could limit the utilization of NOMA in future networks with mass mobile devices. To overcome this challenge, user grouping has been widely studied, especially for the mmWave networks where users have strongly correlated channels. For the mmWave NOMA, the highly directional feature of mmWave transmission increases the correlation of channels, which encourages the implementation of NOMA [19]. Since multiple users can share the same mmWave beam via NOMA, the combination of mmWave and NOMA is highly beneficial to support massive connectivity in a dense area [20]. In [21], the sum rate maximization was investigated by Cui *et al.* through jointly optimizing the user scheduling and PA. Zhu *et al.* first proposed a user grouping method based on the channel correlation, and then the hybrid beamforming and PA were jointly optimized to maximize the achievable sum rate for the downlink mmWave-NOMA networks [22]. In [23], the Stackelberg game was formulated by Wang *et al.* to realize the joint user clustering and PA for the downlink mmWave-NOMA system, and two different optimization problems were proposed to satisfy different demands of users. In [24], the sum rate maximization of the downlink mmWave-NOMA network was studied by optimizing user clustering and PA, where an expectation maximization based algorithm was developed to group the users for both the fixed and dynamic user scenarios.

From the above-mentioned works [21]–[24], it is obvious that all of them intend to optimize the throughput but ignore the security concerns. To the best of our knowledge, there are only a few works discussing the security issue in mmWave-NOMA networks [25]–[27]. In [25], Zhao *et al.* maximized the secure energy efficiency of mmWave NOMA system by optimizing the PA based on a well-designed hybrid beamforming. Huang *et al.* proposed a user pairing scheme based on minimal angle-difference and designed two beaforming methods to enhance the security of mmWave-NOMA networks, with the secrecy outage probability (SOP) analyzed [26]. In [27], a novel framework was developed by Sun *et al.* where the UAV serves the secure user and energy-constrained user simultaneously using mmWave-NOMA.

In this work, we consider the secure transmission in the mmWave-NOMA network with massive connectivity, where a novel hybrid grouping of secure and common users is presented, and the user grouping and PA are jointly optimized to meet the heterogeneous services of users. To the best of our knowledge, very few research works have discussed the joint user grouping and PA optimization problem for mmWave-NOMA networks with security concerns. The key motivations and contributions of this work are summarized as follows.

- For mmWave-NOMA networks with massive connection, user grouping can effectively mitigate the hardware complexity of SIC and support more users. Although the severe intra-group and inter-group interference therein need to be carefully managed, it can be also used to enhance the security as long as the attenuation caused by the interference at the eavesdropper is more serious than that at the secure users. Motivated by this, user grouping and PA are jointly designed for the mmWave-NOMA system to reduce the SIC complexity and guarantee the security and reliability by interference management.
- Firstly, by leveraging the heterogeneous security services in NOMA networks, we present a mixed user grouping framework where each group consists of one secure user (SU) and multiple common users (CUs). To safeguard the SUs, hybrid beamforming towards SUs is designed for each group according to the CSI of SUs.
- To control the information leakage of SUs and improve the transmission capacity of CUs, the sum rate of CUs is maximized with the rate requirement of SUs satisfied, through optimizing the user grouping and PA. The original problem is split into two subproblems of user grouping and power optimization since it belongs to intractable mixed integer nonlinear programming.
- Subsequently, a hybrid SU-CU grouping algorithm based on channel correlations and matching theory is proposed to handle the user grouping. Afterwards, a low-complexity algorithm based on successive convex approximation (SCA) is designed to solve the power optimization, where Lagrangian dual decomposition is adopted to solve the convex PA problem after conversion.

The arrangement of the remaining sections is given as follows. In Section II, the system model is presented, with the beamforming design and problem formulation described in Section III. The hybrid SUs-CUs grouping algorithm is proposed in Section IV, and Section V exploits the SCA-based method to solve the power optimization problem. In Section VI, simulation results are provided, and the conclusions are drawn in Section VII.

Notation: The Hermitian transpose of matrix $\mathbf{A}$ is set as $\mathbf{A}^{\dagger}$. $\|\mathbf{a}\|$ and $\|\mathbf{A}\|$ are the Euclidean norm of vector $\mathbf{a}$ and the Frobenius norm of matrix $\mathbf{A}$, respectively. $\mathbb{C}^{M\times N}$ denotes the space of complex $M \times N$ matrices. $\mathcal{CN}(\mathbf{a}, \mathbf{A})$ means the complex Gaussian distribution with mean $\mathbf{a}$ and covariance matrix $\mathbf{A}$. $|A|$ denotes the cardinality of set A. $A\backslash\{x\}$ means the element x is removed from the set A.

II. System Model

Consider a downlink mmWave NOMA system with one BS, K legitimate users including SUs and CUs, and L eavesdrop-

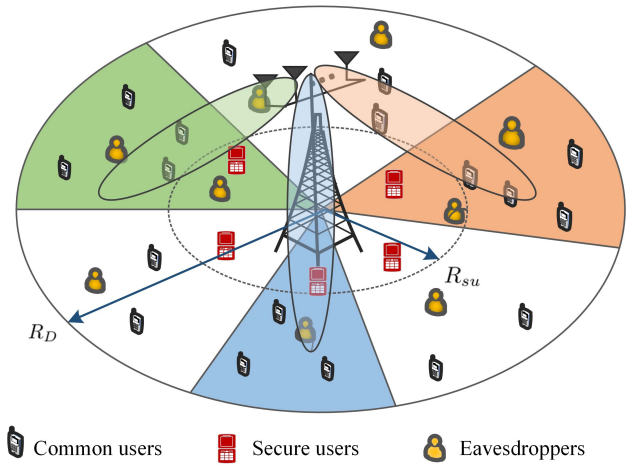


Fig. 1. Illustration of user grouping involving secure and common users in the mmWave-NOMA system with multiple eavesdroppers, where R_{su} and R_D are the radius of the SUs and CUs region, respectively. Security-oriented beamforming is designed for each group.

pers, as shown in Fig. 1. The BS is located at the center of a disc $\mathcal{D}$ with radius R_D. N_t antennas and N_{RF} RF chains are equipped at the BS, and all the other single-antenna nodes are randomly deployed in $\mathcal{D}$. In this paper, we consider an overloaded case that $K > N_{RF}$, and user grouping is applied to support massive connectivity, i.e., M groups are formed to accommodate as many users as possible with $M = N_{RF}$. Denote the set of all the groups as $\mathcal{Q} = \{Q_1, Q_2, \cdots, Q_M\}$ with the set of group index defined as $\mathcal{M} = \{1, 2, \cdots, M\}$, and $|Q_m|$ is the number of users in the mth group. One SU and multiple CUs coexist in each group, and we assume that the strongest user is the SU due to the fact that the security of stronger users with higher channel gains in NOMA networks is intrinsically superior to that of weaker users. The eavesdroppers are only interested in the confidential messages from SUs. In addition, we assume that perfect CSI of all the legitimate users is available at the BS. Without loss of generality, assume that K_{su} SUs are randomly located in the region with radius R_{su}, $R_{su} \leq R_D$, while K_{cu} CUs are randomly placed in the ring area with distance $R_{su} \leq d \leq R_D$. For simplicity, we define the user set of the mth group as $\mathcal{U}_m = \{su, 2, 3, \ldots, |Q_m|\}$, and the set of all the legitimate users can be expressed as

$$\mathcal{K} = \mathcal{K}_{su} \bigcup \mathcal{K}_{cu} = \{1, \cdots, K_{su}\} \bigcup \{1, \cdots, K_{cu}\}. \quad (1)$$

In the mmWave-NOMA network, the BS generates M beams to serve M groups. Thus, the superposed signal transmitted by the BS can be denoted as

$$\mathbf{s} = \sum_{m=1}^{M} \mathbf{w}_m \left(\sum_{j=1}^{|Q_m|} \sqrt{\beta_j^{[m]}} s_j^{[m]} \right), \quad (2)$$

where $\mathbf{w}_m = \mathbf{P}_{RF} \mathbf{p}_{BB}^{[m]}$ is the hybrid precoding vector of group m, and $\mathbf{P}_{RF}$ and $\mathbf{p}_{BB}^{[m]}$ are the analog and digital precoding, respectively. $\beta_j^{[m]}$ denotes the allocated power of user j in the mth group, and $s_j^{[m]}$ is the corresponding information symbol with $\left| s_j^{[m]} \right|^2 = 1$. Considering the limited scattering of mmWave transmission caused by the high free-space path loss, the following geometric channel model with N scatterers is adopted [28],

$$\mathbf{h}_j^{[m]} = \sqrt{\frac{N_t \rho_j^{[m]}}{N}} \sum_{n=1}^{N} g_{j,n}^{[m]} \mathbf{a}\left(\phi_{j,n}^{[m]}\right), \quad (3)$$

where $\mathbf{h}_j^{[m]} \in \mathbb{C}^{N_t \times 1}$. N denotes the number of paths for each user. $\rho_j^{[m]} = \rho_0 \left(d_j^{[m]}\right)^{-\alpha}$, in which ρ_0 is the frequency dependent path loss at reference distance $d_0 = 1$, and $d_j^{[m]}$ is the distance from the BS to user j in the mth group, α is the path-loss exponent with $\alpha = \alpha_{LOS}$ and $\alpha = \alpha_{NLOS}$ for the line-of-sight (LOS) and non-line-of-sight (NLOS) links, respectively. $n = 1$ represents the LOS link in this work. $g_{j,n}^{[m]} \sim \mathcal{CN}(0, 1)$ is the complex channel gain of the nth path of user j. Taking a uniform linear array into account, the antenna array response vector can be given by

$$\mathbf{a}\left(\phi_{j,n}^{[m]}\right) = \frac{1}{\sqrt{N_t}} \left[1, e^{i\pi\phi_{j,n}^{[m]}}, \cdots, e^{i\pi(N_t-1)\phi_{j,n}^{[m]}}\right]^{\dagger}, \quad (4)$$

in which $\phi_{j,n}^{[m]}$ is the normalized direction of the nth path for user j, i.e.,

$$\phi_{j,n}^{[m]} = \frac{2d \sin \theta_{j,n}^{[m]}}{\lambda}, \quad (5)$$

where $\theta_{j,n}^{[m]}$ is the angle of departure of user j. d and λ separately denote the antenna spacing and signal wavelength. Without loss of generality, we assume that $d = \lambda/2$.

Then, the received signal of the jth user in the mth group can be expressed as

$$\begin{aligned} y_j^{[m]} = & \underbrace{\mathbf{h}_j^{[m]\dagger} \mathbf{w}_m \sqrt{\beta_j^{[m]}} s_j^{[m]}}_{\text{desired signal}} \\ & + \underbrace{\mathbf{h}_j^{[m]\dagger} \mathbf{w}_m \sum_{i=1, i\neq j}^{|Q_m|} \sqrt{\beta_i^{[m]}} s_i^{[m]}}_{\text{intra-group interference}} \\ & + \underbrace{\mathbf{h}_j^{[m]\dagger} \sum_{n=1, n\neq m}^{M} \sum_{k=1}^{|Q_n|} \mathbf{w}_n \sqrt{\beta_k^{[n]}} s_k^{[n]}}_{\text{inter-group interference}} + n_j^{[m]}, \end{aligned} \quad (6)$$

where $n_j^{[m]} \sim \mathcal{CN}(0, \sigma^2)$ denotes the additive white Gaussian noise (AWGN) at user j. Assume that the SIC order is as per the descending order of the index of users, i.e., the message of the user $|Q_m|$ is first decoded in the mth group. Thus, the received signal-to-interference-plus-noise ratio (SINR) of the jth user in the mth group can be denoted as

$$\gamma_{j,j}^{[m]} = \begin{cases} \dfrac{\beta_j^{[m]} \left| \mathbf{h}_j^{[m]\dagger} \mathbf{w}_m \right|^2}{I_{intra\_j}^{[m]} + I_{inter\_j}^{[m]} + \sigma^2}, j = 2, \cdots, |Q_m|, \\ \dfrac{\beta_j^{[m]} \left| \mathbf{h}_j^{[m]\dagger} \mathbf{w}_m \right|^2}{I_{inter\_j}^{[m]} + \sigma^2}, j = 1, \end{cases} \quad (7)$$

where the intra-group and inter-group interference can be respectively denoted as

$$\begin{aligned} I_{intra\_j}^{[m]} &= \left|\mathbf{h}_j^{[m]\dagger}\mathbf{w}_m\right|^2 \sum\nolimits_{k=1}^{j-1} \beta_k^{[m]}, \\ I_{inter\_j}^{[m]} &= \sum\nolimits_{n=1,n\neq m}^{M} \left|\mathbf{h}_j^{[m]\dagger}\mathbf{w}_n\right|^2 \sum\nolimits_{k=1}^{|Q_n|} \beta_k^{[n]}. \end{aligned} \tag{8}$$

Accordingly, the received SINR of the jth user decoded at the ith user in the mth group is given by

$$\gamma_{i,j}^{[m]} = \frac{\beta_j^{[m]} \left|\mathbf{h}_i^{[m]\dagger}\mathbf{w}_m\right|^2}{I_{intra\_i}^{[m]} + I_{inter\_i}^{[m]} + \sigma^2}, 1 \leq i \leq j-1, \tag{9}$$

in which $I_{intra\_i}^{[m]} = |\mathbf{h}_i^{[m]\dagger}\mathbf{w}_m|^2 \sum_{k=1}^{j-1} \beta_k^{[m]}$ and $I_{inter\_i}^{[m]} = \sum_{n=1,n\neq m}^{M} |\mathbf{h}_i^{[m]\dagger}\mathbf{w}_n|^2 \sum_{k=1}^{|Q_n|} \beta_k^{[n]}$. To guarantee the success of SIC, the achievable received SINR of user j in the mth group can be denoted as

$$\gamma_j^{[m]} = \begin{cases} \min\limits_{1\leq i\leq j} \left\{\gamma_{i,j}^{[m]}\right\}, \ j = 2, \cdots, |Q_m|, \\ \gamma_{j,j}^{[m]}, \quad j = 1. \end{cases} \tag{10}$$

Based on (10), the intra-group interference caused by user j at the ith user can be removed when $\gamma_j^{[m]}$ meets its minimum SINR requirement, i.e., $\gamma_j^{[m]} \geq \bar{\gamma}_j^{[m]}$.

III. Beamforming and Problem Formulation

To guarantee the security, the secure beamforming towards SUs is designed for each group. Then, the optimization problem is formulated to protect the secrecy of SUs and enhance the transmission capability of CUs by jointly optimizing the user grouping and PA.

A. Security-Oriented Beamforming

We assume that multiple eavesdroppers are randomly distributed to surround the SU and overhear the confidential messages of the SU. As mentioned above, each group has one SU and multiple CUs and the SU is regarded as the strongest one that is the last to be decoded, due to the following two reasons. First, the strongest user can achieve the best performance due to its dominant channel quality and the elimination of intra-group interference. Second, the risk of leaking confidential message is degraded since the lowest power is allocated to the strongest user. Thus, regrading the strongest user as the secure one is helpful to increase the information security.

To make sure that the SU is the strongest one, the security-oriented beamforming is proposed for each group as per the CSI of SU, i.e., a two-stage hybrid precoding method towards each SU is adopted considering the limited number of RF chains [29]. Explicitly, analog precoding is first designed to maximize the antenna gains of SUs. Taking the practical limitation of phase shifts into account, $B-$bit quantized phases are utilized to calculate the analog precoding. In this way, all the elements of the analog precoding vector $\mathbf{P}_{RF}$ should be involved in the following set as

$$\frac{1}{\sqrt{N_t}} \left\{ e^{i\frac{2\pi b}{2^B}}, \ b = 0, 1, \cdots, 2^B - 1 \right\}.$$

In order to boost the signal strength of the SU, the nth element can be designed as

$$\mathbf{p}_{RF}^{[m]}(n) = \frac{1}{\sqrt{N_t}} e^{i\frac{2\pi\hat{b}}{2^B}}, n = 1, \cdots, N_t, \tag{11}$$

in which $\mathbf{p}_{RF}^{[m]} \in \mathbb{C}^{N_t \times 1}$ and

$$\hat{b} = \underset{b\in\{0,1,\cdots,2^B-1\}}{\arg\min} \left| \mathrm{angle}\left(\mathbf{h}_{\mathrm{su}}^{[\mathrm{m}]}(\mathrm{n})\right) - \frac{2\pi \mathrm{b}}{2^{\mathrm{B}}} \right|.$$

Thus, the RF precoding matrix for all the groups is $\mathbf{P}_{RF} = [\mathbf{p}_{RF}^{[1]}, \mathbf{p}_{RF}^{[2]}, \cdots, \mathbf{p}_{RF}^{[M]}] \in \mathbb{C}^{N_t \times M}$. Based on the RF precoding, zero-forcing (ZF) based digital precoding is further used to mitigate the inter-group interference at the SUs, which means that the digital precoding matrix should be

$$\mathbf{P}_{BB} = \widetilde{\mathbf{H}}^{\dagger} \left(\widetilde{\mathbf{H}}\widetilde{\mathbf{H}}^{\dagger}\right)^{-1}, \tag{12}$$

where $\widetilde{\mathbf{H}} = \left[\tilde{\mathbf{h}}_{su}^{[1]\dagger}, \tilde{\mathbf{h}}_{su}^{[2]\dagger}, \cdots, \tilde{\mathbf{h}}_{su}^{[M]\dagger}\right]^{\dagger} \in \mathbb{C}^{M\times M}$, $\tilde{\mathbf{h}}_{su}^{[m]} = \mathbf{h}_{su}^{[m]\dagger}\mathbf{P}_{RF}$ with the dimension of $1 \times M$. The normalized ZF precoding vector for the mth group can be denoted as

$$\mathbf{p}_{BB}^{[m]} = \frac{\mathbf{P}_{BB}^{[m]}}{\left\|\mathbf{P}_{RF}\mathbf{P}_{BB}^{[m]}\right\|}. \tag{13}$$

As a result, the hybrid precoding vector towards the SU in group m can be obtained as

$$\mathbf{w}_{su}^{[m]} = \mathbf{P}_{RF}\mathbf{p}_{BB}^{[m]}, \tag{14}$$

and all the users in the group share the same beam, i.e., $\mathbf{w}_m = \mathbf{w}_{su}^{[m]}$.

When the precoding works, the SUs are free of interference. Hence, the received SNR at the SU in the mth group can be expressed as

$$\gamma_{su}^{[m]} = \frac{\beta_{su}^{[m]} \left|\mathbf{h}_{su}^{[m]\dagger}\mathbf{w}_m\right|^2}{\sigma^2}. \tag{15}$$

However, as for passive eavesdroppers, it is difficult to implement SIC due to the lack of the ordering for legitimate channel gains. Therefore, both the intra-group and inter-group interference will disturb the eavesdropping, i.e., the intercepted SINR of SU at the lth eavesdropper [1] in the mth group can be expressed as

$$\gamma_{el}^{[m]} = \frac{\beta_{su}^{[m]} \left|\mathbf{h}_{el}^{\dagger}\mathbf{w}_m\right|^2}{I_{intraE}^{[m]} + I_{interE}^{[m]} + \sigma^2}, \tag{16}$$

where

$$\begin{aligned} I_{intraE}^{[m]} &= \left|\mathbf{h}_{el}^{\dagger}\mathbf{w}_m\right|^2 \sum\nolimits_{k=2}^{|Q_m|} \beta_k^{[m]}, \\ I_{interE}^{[m]} &= \sum\nolimits_{n=1,n\neq m}^{M} \left|\mathbf{h}_{el}^{\dagger}\mathbf{w}_n\right|^2 \sum\nolimits_{k=1}^{|Q_n|} \beta_k^{[n]}. \end{aligned} \tag{17}$$

[1] In this work, we consider the passive eavesdropping case where eavesdroppers just keep listening without transmitting any information. In this case, it is hard to obtain the eavesdroppers' CSI, and the secrecy rate cannot be calculated at the BS. Note that the expressions of the eavesdropping SINR in (16) and the achievable secrecy rate in (18) are only utilized to estimate the possible eavesdropping and secure capacity achieved in this work, respectively. This can reflect the performance of the proposed scheme.

Considering the worst-case intercepted SINR among L eavesdroppers, the achievable secrecy rate of the SU in the mth group can be defined as

$$R_s^{[m]} = \left[\log_2\left(1+\gamma_{su}^{[m]}\right) - \log_2\left(1+\max_{1\le l\le L}\left\{\gamma_{el}^{[m]}\right\}\right)\right]^+, \quad (18)$$

where $[x]^+ \triangleq \max(x,0)$. From (18), it is obvious that the SU in each group can achieve interference-free transmission, whereas the eavesdropping performance can be severely degraded by both the intra-group and inter-group interference. Thus, the privacy of SUs can be protected via precoding.

B. Problem Formulation

We consider the case when the eavesdropping CSI is unknown at the BS. In this case, the transmit power of SU should be carefully controlled to mitigate the secrecy information leakage. On the other hand, to guarantee the transmission performance of CUs, the user scheduling and PA need to be well designed. Thus, the optimization problem can be formulated as

$$\max_{\boldsymbol{\zeta},\boldsymbol{\beta}} \sum_{m\in\mathcal{M}}\sum_{j\in\mathcal{K}\backslash\mathcal{K}_{su}} \zeta_j^{[m]}\log_2\left(1+\gamma_j^{[m]}\right) \quad (19a)$$

$$s.t. \sum_{su\in\mathcal{K}_{su}} \zeta_{su}^{[m]}\log_2\left(1+\gamma_{su}^{[m]}\right) \ge r_{su}^{[m]},\ m\in\mathcal{M}, \quad (19b)$$

$$\sum_{j\in\mathcal{K}\backslash\mathcal{K}_{su}} \zeta_j^{[m]}\log_2\left(1+\gamma_j^{[m]}\right) \ge r_j^{[m]}, \quad (19c)$$

$$\sum_{m\in\mathcal{M}}\left(\beta_{su}^{[m]} + \sum_{q\in\mathcal{U}_m\backslash\{su\}}\beta_q^{[m]}\right) \le P, \quad (19d)$$

$$\sum_{m\in\mathcal{M}} \zeta_j^{[m]} \le 1,\ j\in\mathcal{K}, \quad (19e)$$

$$\sum_{j\in\mathcal{K}} \zeta_j^{[m]} = |Q_m|,\ m\in\mathcal{M}. \quad (19f)$$

In (19), given the different service requirements of SU and CUs, we intend to maximize the sum rate of CUs while safeguarding the secure transmission of SUs by jointly optimizing the user scheduling and PA. The vectors $\boldsymbol{\zeta} = \left\{\zeta_j^{[m]}\in\{0,1\}|j\in\mathcal{K},m\in\mathcal{M}\right\}$ and $\boldsymbol{\beta} = \left\{\beta_u^{[m]}|u\in\mathcal{U}_m,m\in\mathcal{M}\right\}$ denote the set of user scheduling index and PA factors, respectively. The jth user is scheduled on the mth group when $\zeta_j^{[m]}=1$, $r_{su}^{[m]}$ and $r_j^{[m]}$ are the data rate thresholds of the SU and CUs, respectively, and P is the total transmit power of BS.

In particular, the constraint (19b) is the rate requirement of SUs, which aims at satisfying the quality of service (QoS) of each SU and leaking its information as little as possible, (19c) is the minimum data rate constraint of CUs, (19d) denotes the power consumption limitation, (19e) indicates that at most one user is scheduled in a single beam, and (19f) ensures that each beam is multiplexed by $|Q_m|$ users. Note that the problem (19) belongs to the mixed integer nonlinear programming involving both the continuous and combinational integer variables, which is intractable to solve directly. Thus, in the following sections, we decouple the original problem into two sub-problems, and utilize the matching theory and SCA to obtain its suboptimal solution.

IV. Hybrid Grouping Algorithm

To solve the original problem, (19) can be decoupled into two subproblems, user grouping and power optimization, to be solved separately. We first consider the hybrid SU-CU grouping problem with fixed PA coefficients. According to [23], the decoding order can be adjusted to satisfy the condition of SIC when the PA coefficients are determined, i.e., $\gamma_{i,j}^{[m]} > \gamma_{j,j}^{[m]}, 1\le i\le j-1$ should be always met through optimizing the decoding order. In this way, $\gamma_j^{[m]} = \gamma_{j,j}^{[m]}$, which means that the original problem can be rewritten as

$$\max_{\boldsymbol{\zeta}} \sum_{m\in\mathcal{M}}\sum_{j\in\mathcal{K}\backslash\mathcal{K}_{su}} \zeta_j^{[m]}\log_2\left(1+\gamma_{j,j}^{[m]}\right) \quad (20a)$$

$$s.t. \sum_{m\in\mathcal{M}} \zeta_j^{[m]} \le 1,\ j\in\mathcal{K} \quad (20b)$$

$$\sum_{j\in\mathcal{K}} \zeta_j^{[m]} = |Q_m|,\ m\in\mathcal{M}. \quad (20c)$$

From (20), it can be seen that the effect of SUs on the objective function is only through the intra-group and inter-group interference they caused to the CUs. This means that lower power allocated to the SUs is beneficial to maximize the sum rate of CUs. Thus, the cluster-head selection algorithm in [30] is utilized to reinforce the strength of channel gains of SUs as well as suppressing the inter-group interference.

Specifically, the user with the largest channel gain is selected as the first SU, i.e.,

$$SU_1 = \arg\max_{k\in\mathcal{K}_{su}} \|\mathbf{h}_k\|^2. \quad (21)$$

Then, the channel correlation between SU_1 and other remaining users is calculated and the users whose channel correlation with SU_1 is smaller than the threshold ε are chosen as the candidates of other SUs. Among the candidates, the user that has the highest channel gain is chosen as the second SU. Next, the channel correlation between the second SU and the remaining candidates are measured, and the candidates of SUs are also updated as per the threshold ε. The third SU is selected using the same method as the second SU. Repeat the above procedure until all the SUs over M beams are determined. Note that, during the selection, a small increment will be added on the threshold if the set of candidates is null and the number of SUs is less than M.

After the SUs of all the groups are settled, the problem (20) is reduced to a sum rate maximization problem by optimizing the user scheduling as

$$\max_{\boldsymbol{\zeta}} \sum_{m\in\mathcal{M}}\sum_{j\in\mathcal{K}\backslash\mathcal{K}_{su}} \zeta_j^{[m]}\log_2\left(1+\gamma_{j,j}^{[m]}\right) \quad (22a)$$

$$s.t. \sum_{m\in\mathcal{M}} \zeta_j^{[m]} \le 1,\ j\in\mathcal{K}\backslash\{su\} \quad (22b)$$

$$\sum_{j\in\mathcal{K}\backslash\mathcal{K}_{su}} \zeta_j^{[m]} = |Q_m|-1,\ m\in\mathcal{M}. \quad (22c)$$

To solve this problem, the matching theory can be applied. As discussed above, each beam can be shared by multiple CUs while each CU is assigned to at most one group. Thus, the relationship between the groups and CUs can be mapped as a many-to-one matching problem, which will be elaborated next.

A. Many-to-one Matching

Before giving the notion of many-to-one matching, we first define ψ as the matching symbol, i.e., $\psi(k)$ means the beam assigned to user k, and $\psi(m)$ corresponds to all the CUs that are scheduled in group m. Based on this, the many-to-one matching can be defined as follows.

Definition 1: Considering two disjoint sets of participants $\mathcal{K}_{cu}$ and $\mathcal{M}$, a many-to-one matching ψ is a correspondence from all the subsets of $\mathcal{K}_{cu}$ into the group set $\mathcal{M}$ that satisfies

1) $\psi(m) \subseteq \mathcal{K}$ with $|\psi(m)| = |Q_m|$-1 for all $m \in \mathcal{M}$;
2) $\psi(k) \subseteq \mathcal{M}$ with $|\psi(k)| = 1$ for all $k \in \mathcal{K}_{cu}$;
3) $k \in \psi(m) \Longleftrightarrow \psi(k) = m$ for all $m \in \mathcal{M}$ and $k \in \mathcal{K}_{cu}$.

Conditions 1) and 2) represent quota constraints of the mth group and the kth CU. Condition 3) depicts the matching relationship between group m and CU_k.

Before solving the matching problem, the utility functions and preference profile of CUs and groups need to be predetermined. In this work, both of them are defined based on their achievable transmission rate. Specifically, the utility function of the ith CU in the mth group under the matching ψ can be given as

$$\boldsymbol{U}_i(\psi) = \log_2\left(1+\gamma_{i,i}^{[m]}\right), i \in \mathcal{U}_m \backslash \{su\}. \tag{23}$$

The utility function of the mth group can be also modeled as

$$\boldsymbol{U}_m(\psi) = \sum_{i \in \psi(m)} \log_2\left(1+\gamma_{i,i}^{[m]}\right), m \in \mathcal{M}. \tag{24}$$

According to the above utility functions, the preference lists (PLs) between the CUs and groups can be established as

$$\begin{aligned} \mathcal{P}_{cu} &= [\mathcal{P}_{cu}(1), \cdots, \mathcal{P}_{cu}(i), \cdots, \mathcal{P}_{cu}(K_{cu})], \\ \mathcal{P}_g &= [\mathcal{P}_g(1), \cdots, \mathcal{P}_g(m), \cdots, \mathcal{P}_g(M)], \end{aligned} \tag{25}$$

where $\mathcal{P}_{cu}(i)$ and $\mathcal{P}_g(m)$ are the PL of user i and group m with regard to the group set $\mathcal{M}$ and the user set $\mathcal{K}_{cu}$, respectively. To make it clear, we use $\succ$ to symbolize the preference relationship, i.e., $\succ_i, (i \in \mathcal{K}_{cu})$ denotes the preference order of the ith user over the group set $\mathcal{M}$, and the preference relationship between the given group m and the user set $\mathcal{K}$ is defined as $\succ_m$. For a pointed user i, we consider two different groups m and $\bar{m}$ that belong to the set $\mathcal{M}$ and relate to the matching ψ and $\bar{\psi}$, respectively. We define

$$(m, \psi) \succ_i (\bar{m}, \bar{\psi}) \Longleftrightarrow \boldsymbol{U}_i(\psi) > \boldsymbol{U}_i(\bar{\psi}), \tag{26}$$

which means that user i is more preferable to be allocated to group m under the matching ψ than group $\bar{m}$ with matching $\bar{\psi}$ due to the higher transmission rate of user i achieved in group m. As well, for any two matched user sets $\mathcal{U}_j$ and $\mathcal{U}_k$ with their corresponding matching ψ and $\bar{\psi}$, we introduce

$$(\mathcal{U}_j, \psi) \succ_m (\mathcal{U}_k, \bar{\psi}) \Longleftrightarrow \boldsymbol{U}_m(\psi) > \boldsymbol{U}_m(\bar{\psi}). \tag{27}$$

In this case, group m prefers to choose the user set $\mathcal{U}_j$ for better performance.

Remark 1: From the above analysis, it is evident that there exists interdependency between the utility functions and PLs of these two distinct participant sets, which implies that the problem (22) is a two-sided matching problem with externalities. Specifically, the externalities can be explained as that the achievable rate of each CU not only depends on the interference from the internal users, but also relates to the external interference caused by the users of other groups. Similarly, the throughput of each group is not just up to the users in this group, but also affected by the users from other groups.

B. Two-Sided Swap Matching

Due to the externalities of the matching problem (22), the traditional matching algorithms are not feasible to obtain a stable matching. Thus, the principle of two-sided exchange-stability (TES) in [31] is adopted to find the stable matching of the problem (16), which aims at searching the swap blocking pair through exchanging any two users over different groups constantly and end to a stable matching. To better understand the swap matching method, we first introduce the concept of swap matching and swap blocking pair.

Definition 2: For any two distinct CUs, i and j, the swap matching can be defined as

$$\psi_i^j = \{\psi \backslash \{(i,m),(j,n)\} \cup \{(j,m),(i,n)\}\}, \tag{28}$$

where $\psi(i) = m$ and $\psi(j) = n$.

Definition 2 reveals that every swap matching operation only involves two CUs and two different groups they belong to, whereas has no effect on the profits of all other groups and their supported CUs.

Definition 3: According to the concept of blocking pair [32], the swap blocking pair in this work can be depicted as follows. Given one CU pair (i, j) with $\psi(i) = m$ and $\psi(j) = n$, both of them have the incentive to deviate from their current groups and establish a new matching with each other's group, i.e., the previous combination $\{(i,m),(j,n)\}$ is changed to $\{(i,n),(j,m)\}$. As such, the primal CU pair is called a swap blocking pair.

Based on this, how to judge whether one CU pair is a swap blocking one, i.e., the conditions whether the swap matching is permitted, is given in the succeeding definition.

Definition 4: The swap matching can be approved if and only if one CU pair satisfies

$$\begin{aligned} &\forall x \in \{i, j, \psi(i), \psi(j)\},\ \boldsymbol{U}_x(\psi_i^j) \geq \boldsymbol{U}_x(\psi), \\ &\exists x \in \{i, j, \psi(i), \psi(j)\},\ \boldsymbol{U}_x(\psi_i^j) > \boldsymbol{U}_x(\psi), \end{aligned}$$

where $\boldsymbol{U}_x$ denotes the utility of the participant x involved in the swap matching, and the CU pair (i, j) is deemed as a swap blocking pair when the swap matching is allowed.

From Definition 4, it can be seen that a pair of CUs on different groups can be exchanged when the utility of all the members considered in the swap matching will not decrease, and at least the achievable rate of one member will increase. Following the concept of swap matching, every two users

belong to different beams are first formed as a user pair. After performing a sequence of swap matching operations based on Definition 4, the matching will finally approach a novel kind of stability, i.e., two-sided exchange-stability. In particular, the matching is two-sided exchange-stable when no user pair can be determined as a swap blocking pair, i.e., for any matched user pair, the swap matching is no longer approved.

C. Hybrid SU-CU Grouping Algorithm

Based on the above analysis, the hybrid SU-CU grouping process can be summarized as Algorithm 1. First, the SUs are chosen as per the cluster-head selection algorithm in [30] to boost the signal strength of SUs as well as reduce the intra-group and inter-group interference of CUs, which is helpful to maximize the objective function of (20). After that, the CU grouping problem can be solved via the matching theory based on TES. Particularly, the CU grouping process can be divided into two steps, i.e., initialization and swap matching. In the initialization, the channel correlation between the CUs and the selected SUs are calculated as

$$\mathcal{C}_k^{[m]} = \frac{\left|\tilde{\mathbf{h}}_{cu}^{k}\tilde{\mathbf{h}}_{su}^{[m]\dagger}\right|}{\left\|\tilde{\mathbf{h}}_{cu}^{k}\right\|^2\left\|\tilde{\mathbf{h}}_{su}^{[m]}\right\|^2}, k \in \mathcal{K}_{cu}, m \in \mathcal{M}. \tag{29}$$

Then, the Gale-Shapley algorithm based on the channel correlation is utilized to obtain the initial matching set. Based on this, the swap matching step aims at approaching a stable matching between the CUs and the groups through searching the swap blocking pair and updating the matching set iteratively.

Remark 2: In Algorithm 1, the two-sided matching can always achieve the stable status due to the finite number of swap matching operations from the limited size of the matched set.

Remark 3: All the local optimal solutions to the problem (22) correspond to a TES, yet not all the TES matchings are local optimal solutions. From Definition 4, it can be deduced that the CU_j over the mth group will reject the exchanging when its transmission rate decreases, though the swap operation may lead to a higher sum rate of all the users. In this case, if the swap matching is imposed, the sum rate of all the users will increase but at the cost of one-sided matching stability. Thus, not all the TES matchings can converge to local optimal solutions, and the rate loss can be deemed as the expense of ensuring the fairness of resource allocation.

Remark 4: Although Algorithm 1 cannot obtain the global optima, its complexity can be significantly reduced compared with the exhaustive method. For clarity, the complexity of the SU grouping is $\mathcal{O}(MK_{su}^2)$ [30]. The maximum number of proposals in the initialization step of the CU grouping can be deduced as $(K_{cu}M)$, and the number of swap actions in each iteration is calculated as $\frac{1}{2}(|Q_m|-1)^2M(M-1)$. Therefore, Algorithm 1 has polynomial complexity.

Algorithm 1 Hybrid SU-CU Grouping Algorithm

SU Grouping Method: Using the cluster-head selection algorithm to select SUs that have small channel similarity and high channel gains.

CU Grouping Method: ***Initialization***

Calculate the correlation of equivalent channels between all the unmatched CUs and the chosen SUs. Initialize the PLs $\mathcal{P}_{cu}$ and $\mathcal{P}_g$ as per the descending order of the correlation. Set the user set accepted by the mth group $\mathcal{S}_m = \emptyset$ and the unmatched user set $\mathcal{D}_u = \mathcal{K}_{cu}$. Set $n = 1$.

while $(|\mathcal{S}_m| \neq |Q_m| - 1$ *or* $\mathcal{P}_{cu} \neq \emptyset)$ **do**

  In the nth round, all the unassigned CUs propose to the groups that are currently most relevant to them and remove the according groups from their PLs.

  if *the proposals are accepted* **then**

    Gather the accepted user set of group m as $\bar{\mathcal{S}}_m = \mathcal{S}_m^{(n-1)} \bigcup \{\bar{u}_1, \cdots, \bar{u}_s\}$, and select the top $(|Q_m| - 1)$ most relevant users based on $\mathcal{P}_g(m)$, i.e., the updated accepted user set of group m is $\mathcal{S}_m^{(n)} = \{u_1, \cdots, u_{|Q_m|-1}\}$.

  else

    Merge the rejected users into the unassigned user set, i.e., renew the unmatched user set as $\mathcal{D}_u^{(n)} = \mathcal{K}_{cu} \backslash \bigcup_{m \in \mathcal{M}} \{S_m^{(n)}\}$.

  end

  Update: $n = n + 1$.

end

Get the initial matched set as $\mathcal{S} = \{\mathcal{S}_m | m \in \mathcal{M}\}$.

Swap matching

while *there exists a swap blocking pair* **do**

  Searching Step: For the CU $i \in \mathcal{S}$ with $\psi(i) = m$, search the other CU $j \in \mathcal{S} \backslash \mathcal{S}(\psi(m))$ with $\psi(j) = n$. Calculate the achievable rate of all the members before and after exchanging, and determine whether the swap matching is permitted as per Definition 4.

  if (i, j) *is a swap-blocking pair* **then**

    $\psi(i) = n, \psi(j) = m$, update the matched set $\mathcal{S} = \psi_i^j$ and back to the *Searching Step*.

  else

    Remain the original matching and back to the *Searching Step*.

  end

end

Output: The result of stable matching $\mathcal{S}$.

V. Transmit Power Allocation Optimization

According to the obtained user grouping from Algorithm 1, we further optimize the power allocation of all the matched users in this section. The PA problem can be reorganized as

$$\max_{\boldsymbol{\beta}} \sum_{m=1}^{M}\sum_{j=2}^{|Q_m|} \log_2\left(1+\gamma_j^{[m]}\right) \tag{30a}$$

$$s.t.\ \log_2\left(1+\gamma_j^{[m]}\right) \geq r_j^{[m]}, j \in \mathcal{U}_m \backslash \{su\}, \tag{30b}$$

$$\log_2\left(1+\gamma_{su}^{[m]}\right) \geq r_{su}^{[m]},\ m \in \mathcal{M}, \tag{30c}$$

$$\sum_{m=1}^{M}\sum_{j \in \mathcal{U}_m} \beta_j^{[m]} \leq P. \tag{30d}$$

Note that the QoS requirements for SUs (30c) and the power consumption limitation (30d) are convex. However, the problem (30) is still non-convex due to the complicated objective function (30a) and the constraints (30b). To make it solvable, SCA is employed to convert the non-convex problem (30) into a sequence of convex subproblems. Specifically, to handle the non-convex items in the problem (30), we first introduce the following lower bound to the expression of transmission rate [33], i.e., the difference of two logarithmic functions can be transformed by

$$\begin{cases} \ln(1+\gamma) \geq u\ln(\gamma)+v, \\ u=\dfrac{\bar{\gamma}}{1+\bar{\gamma}}, \\ v=\ln(1+\bar{\gamma})-\dfrac{\bar{\gamma}}{1+\bar{\gamma}}\ln\bar{\gamma}, \end{cases} \tag{31}$$

where the inequality is tight when $\gamma=\bar{\gamma}$.

Using (31), the objective function (30a) can be written as

$$\sum_{m=1}^{M}\sum_{j=2}^{|Q_m|}\frac{1}{\ln 2}\left(u_j^{[m]}\ln\left(\gamma_j^{[m]}\right)+v_j^{[m]}\right), \tag{32}$$

which is not concave as well. Nonetheless, by introducing $t_j^{[m]}=\ln(\beta_j^{[m]})$, it can be converted as

$$\begin{aligned} &\ln\left(\gamma_{i,j}^{[m]}\left(\boldsymbol{t}\right)\right)=\ln\left(G_{im}^{[m]}\right)+t_j^{[m]} \\ &-\ln\left(G_{im}^{[m]}\sum_{k=1}^{j-1}e^{t_k^{[m]}}+\sum_{n\neq m}G_{im}^{[n]}\sum_{q=1}^{|Q_n|}e^{t_q^{[n]}}+\sigma^2\right), 1\leq i\leq j, \end{aligned} \tag{33}$$

where $G_{im}^{[m]}=|\mathbf{h}_i^{[m]\dagger}\mathbf{w}_m|^2$. It is obvious that (33) is concave due to the fact that the items in the log-sum-exp form are convex. According to [34], we also know that the minimum of multiple concave functions is still concave. Hence, substituting (33) into (32), the objective function (30a) becomes concave.

As well, the constraint (30b) can be rewritten as

$$\tilde{R}_j^{[m]}\left(\boldsymbol{t}\right)=\frac{1}{\ln 2}\left(u_j^{[m]}\ln\left(\gamma_j^{[m]}\left(\boldsymbol{t}\right)\right)+v_j^{[m]}\right)\geq r_j^{[m]}, \tag{34}$$

which is also transformed to convex.

Using the above transformations, the problem (30) can be reformulated as a convex one, i.e.,

$$\max_{\boldsymbol{t}}\ \sum_{m=1}^{M}\sum_{j=2}^{|Q_m|}\frac{1}{\ln 2}\left(u_j^{[m]}\min_{1\leq i\leq j}\left\{\ln\left(\gamma_{i,j}^{[m]}\right)\right\}+v_j^{[m]}\right) \tag{35a}$$

$$s.t.\ \frac{1}{\ln 2}\left(u_j^{[m]}\min_{1\leq i\leq j}\left\{\ln\left(\gamma_{i,j}^{[m]}\right)\right\}+v_j^{[m]}\right)\geq r_j^{[m]}, \tag{35b}$$

$$t_{su}^{[m]}\geq\ln\left(\frac{(2^{r_{su}^{[m]}}-1)\sigma^2}{G_{su}^{[m]}}\right), \tag{35c}$$

$$\sum_{m=1}^{M}\sum_{j\in\mathcal{U}_m}e^{t_j^{[m]}}\leq P. \tag{35d}$$

In the problem (35), the objective function monotonically decreases with respect to the transmit power of SUs. Thus, the linear constraints (35c) are tightened when the optimal solution is obtained, which means the transmit power of SUs can be derived as

$$\beta_{su}^{[m]}=\frac{\left(2^{r_{su}^{[m]}}-1\right)\sigma^2}{G_{su}^{[m]}}, m\in\mathcal{M}. \tag{36}$$

After that, the problem (35) can be modified as

$$\max_{\tilde{\boldsymbol{t}}}\ \sum_{m=1}^{M}\sum_{j=2}^{|Q_m|}\frac{1}{\ln 2}\left(u_j^{[m]}\min_{1\leq i\leq j}\left\{\ln\left(\gamma_{i,j}^{[m]}\right)\right\}+v_j^{[m]}\right) \tag{37a}$$

$$s.t.\ \frac{1}{\ln 2}\left(u_j^{[m]}\min_{1\leq i\leq j}\left\{\ln\left(\gamma_{i,j}^{[m]}\right)\right\}+v_j^{[m]}\right)\geq r_j^{[m]}, \tag{37b}$$

$$\sum_{m=1}^{M}\sum_{j=2}^{|Q_m|}e^{t_j^{[m]}}\leq\tilde{P}=P-\sum_{m=1}^{M}\beta_{su}^{[m]}, \tag{37c}$$

where $\tilde{\boldsymbol{t}}=\boldsymbol{t}\backslash\boldsymbol{t_{su}}$ and $\boldsymbol{t}_{su}=\left[t_{su}^{[m]}\right]_{m\in\mathcal{M}}$. Note that the problem (37) is convex and easy to solve using existing toolboxes, like CVX. However, in order to reduce the computational complexity, we present a low-complexity algorithm based on Lagrangian dual decomposition to obtain its optimal solution [35]. This means that we can solve the problem (37) by handling its dual problem, and the dual gap becomes zero when strong duality holds. Accordingly, the Lagrangian function related to the problem (37) can be given as

$$\begin{aligned} \mathcal{L}(\tilde{\boldsymbol{t}},\boldsymbol{\vartheta},\lambda)=&\sum_{m=1}^{M}\sum_{j=2}^{|Q_m|}\tilde{R}_j^{[m]}\left(\tilde{\boldsymbol{t}}\right)-\lambda\left(\sum_{m=1}^{M}\sum_{j=2}^{|Q_m|}e^{t_j^{[m]}}-\tilde{P}\right) \\ &+\sum_{m=1}^{M}\sum_{j=2}^{|Q_m|}\vartheta_j^{[m]}\left(\tilde{R}_j^{[m]}\left(\tilde{\boldsymbol{t}}\right)-r_j^{[m]}\right), \end{aligned} \tag{38}$$

where $\boldsymbol{\vartheta}$ and λ are the dual variables corresponding to the constraints (37b) and (37c), respectively. Then, the dual problem associated with the problem (37) can be formulated as

$$\begin{aligned} &\min_{\lambda,\boldsymbol{\vartheta}}\ f\left(\lambda,\boldsymbol{\vartheta}\right) \\ &s.t.\ \lambda\geq 0,\ \boldsymbol{\vartheta}\geq 0, \end{aligned} \tag{39}$$

and the Lagrange dual function is

$$f\left(\lambda,\boldsymbol{\vartheta}\right)=\max_{\tilde{\boldsymbol{t}}}\ \mathcal{L}(\tilde{\boldsymbol{t}},\boldsymbol{\vartheta},\lambda). \tag{40}$$

Based on the dual problem, we can get the optimal solution to the original problem (37) via Proposition 1.

Proposition 1: According to the dual problem (39), the optimal results for the convex problem (37) can be obtained by updating equations (41) (at the top of the next page), (42) and (43) until the iteration converges.

$$\lambda(s+1)=\left[\lambda(s)+\xi_1\left(\sum_{m=1}^{M}\sum_{j=2}^{|Q_m|}\tilde{\beta}_j^{[m]}(s)-\tilde{P}\right)\right]^{+}, \tag{42}$$

$$\vartheta_j^{[m]}(s+1)=\left[\vartheta_j^{[m]}(s)-\xi_2\left(\tilde{R}_j^{[m]}\left(\tilde{\boldsymbol{\beta}}(s)\right)-r_j^{[m]}\right)\right]^{+}, \tag{43}$$

where ξ_1 and ξ_2 denote sufficiently small step sizes, and s is the iteration index.

Proof: Considering the convex optimization theory [34], we can obtain the optimal solution to the original problem by

$$\tilde{\beta}_j^{[m]}(s+1)=\frac{u_{j-1}^{[m]}\left(1+\vartheta_{j-1}^{[m]}\right)}{\lambda\ln 2+\sum\limits_{k=j+1}^{|Q_m|}u_{k-1}^{[m]}\left(1+\vartheta_{k-1}^{[m]}\right)\min\limits_{1\le i\le k}\left\{\frac{\gamma_{i,k}^{[m]}(s)}{\beta_k^{[m]}(s)}\right\}+\sum\limits_{n=1,n\neq m}^{M}\sum\limits_{q=2}^{|Q_n|}u_{q-1}^{[n]}\left(1+\vartheta_{q-1}^{[n]}\right)\min\limits_{2\le i\le q}\left\{\frac{\gamma_{i,q}^{[n]}(s)G_{in}^{[m]}}{G_{in}^{[n]}\beta_q^{[n]}(s)}\right\}}. \quad (41)$$

$$\frac{\partial L}{\partial\tilde{t}_j^{[m]}}=0=\tilde{u}_{j-1}^{[m]}\left(1+\vartheta_{j-1}^{[m]}\right)-e^{\tilde{t}_j^{[m]}}\left(\lambda+\sum_{k=j+1}^{|Q_m|}\tilde{u}_{k-1}^{[m]}\left(1+\vartheta_{k-1}^{[m]}\right)\min_{1\le i\le k}\left\{\frac{\gamma_{i,k}^{[m]}}{e^{t_k^{[m]}}}\right\}+\sum_{n=1,n\neq m}^{M}\sum_{q=2}^{|Q_n|}\tilde{u}_{q-1}^{[n]}\left(1+\vartheta_{q-1}^{[n]}\right)\min_{2\le i\le q}\left\{\frac{\gamma_{i,q}^{[n]}G_{in}^{[m]}}{G_{in}^{[n]}e^{t_q^{[n]}}}\right\}\right), \quad (44)$$

solving its corresponding dual problem. Thus, to obtain the optimal solution to the problem (39), we first solve the dual maximization of $\tilde{\boldsymbol{t}}$, with the dual variables $\boldsymbol{\vartheta}$ and λ fixed. The solution to the maximization problem (40) is unique due to its strict concavity associated with $\tilde{\boldsymbol{t}}$, which can be derived via its first-order derivative with respect to $\tilde{\boldsymbol{t}}$ as (44) at the top of next page. In (44), $\tilde{u}_j^{[m]}=\frac{u_j^{[m]}}{\ln 2}$ and

$$\gamma_{i,k}^{[m]}=\frac{e^{t_k^{[m]}}G_{im}^{[m]}}{\sum\limits_{x=1}^{k-1}e^{t_x^{[m]}}G_{im}^{[m]}+\sum\limits_{n=1,n\neq m}^{M}\sum\limits_{q=1}^{|Q_n|}e^{t_q^{[n]}}G_{im}^{[n]}+\sigma^2}, \quad (45)$$

$$\gamma_{i,q}^{[n]}=\frac{e^{t_q^{[n]}}G_{in}^{[n]}}{\sum\limits_{y=1}^{q-1}e^{t_y^{[n]}}G_{in}^{[n]}+\sum\limits_{z=1,z\neq n}^{M}\sum\limits_{q=1}^{|Q_z|}e^{t_q^{[z]}}G_{in}^{[z]}+\sigma^2}. \quad (46)$$

Replacing $\tilde{t}_j^{[m]}$ with $\tilde{\beta}_j^{[m]}$, we can obtain the fixed-point equation (41) with regard to the PA factors. In the following, the minimization problem (39) can be solved using the gradient-descent method thanks to the differentiable dual function, which means that the dual variables can be updated as (42) and (43).

To solve the problem (37), the dual optimization problem is calculated iteratively in terms of (41), (42) and (43), and the optimal solution can be obtained as the iteration converges. ■

Note that we can only get the optimal values of the problem (37) resorting to Lagrangian dual decomposition when the variables $\boldsymbol{u}$ and $\boldsymbol{v}$ are predefined. Therefore, $\boldsymbol{u}$ and $\boldsymbol{v}$ should be constantly updated to solve the original problem (30) due to the approximations adopted in (31). The iteration process will converge when the increment of the sum rate of SUs is no larger than the maximum tolerance ϵ. Details of the SCA-based algorithm are summarized in Algorithm 2.

Remark 5: Based on conclusions in [33] and [36], Algorithm 2 can converge to a Karush-Kuhn-Tucker (KKT) solution to the original problem (30). Specifically, during per iteration, the value of the objective function becomes larger than or equal to its value in the preceding iteration, which proves that the sum rate of CUs will be monotonically increasing or nondecreasing with the increasing number of iterations. Besides, both the QoS requirements of users and the power consumption limitation constrain the growth of the sum rate of CUs, which ensures that Algorithm 2 is convergent.

Algorithm 2 SCA-based Algorithm

1: Initialization: Set $\boldsymbol{u}_0=\text{ones}(|\text{Q}_\text{m}|-1,\text{M})$, $\boldsymbol{v}_0=\text{zeros}(|\text{Q}_\text{m}|-1,\text{M})$ and $RT=\text{zeros}(|\text{Q}_\text{m}|-1,\text{M})$. Set $\tilde{\boldsymbol{\beta}}=\text{zeros}(|\text{Q}_\text{m}|-1,\text{M})$, $\boldsymbol{\vartheta}=\text{zeros}(|\text{Q}_\text{m}|-1,\text{M})$ and $\lambda=0$. Define the maximum number of iterations as N, and the index of the first iteration is $n=1$.
2: **repeat**
3: **repeat**
4: With $\boldsymbol{u}_n$ and $\boldsymbol{v}_n$ fixed, calculate the PA factors $\tilde{\boldsymbol{\beta}}$ according to (41).
5: Update the vector of the target rate multipliers $\boldsymbol{\vartheta}$ as per (43) until the data rate of all the CUs meets the minimum rate threshold. If $R_j^{[m]}>r_j^{[m]}$, $\vartheta_j^{[m]}=0$. Elsewise, update $\vartheta_j^{[m]}$ using the bisection method to ensure $\left|R_j^{[m]}-r_j^{[m]}\right|\le\bar{\epsilon}$. $\bar{\epsilon}=10^{-5}$.
6: Update the multiplier λ as per (42). If $\sum_{m=1}^{M}\sum_{j=2}^{|Q_m|}\tilde{\beta}_j^{[m]}-\tilde{P}<0$, $\lambda=0$. Otherwise, use the bisection method to renew λ guaranteeing
$$\tilde{P}-\bar{\epsilon}\le\sum_{m=1}^{M}\sum_{j=2}^{|Q_m|}\tilde{\beta}_j^{[m]}\le\tilde{P}.$$
7: **until** $\boldsymbol{\vartheta}$ and λ are convergent.
8: Use the updated PA coefficient to calculate all the CUs' date rate as RT_n.
9: Compute the current SINR as $\bar{\boldsymbol{\gamma}}_n=2^{RT_n}-1$, and update $\boldsymbol{u}_n=\frac{\bar{\boldsymbol{\gamma}}_n}{1+\bar{\boldsymbol{\gamma}}_n}$ and $\boldsymbol{v}_n=\ln(1+\bar{\boldsymbol{\gamma}}_n)-\frac{\bar{\boldsymbol{\gamma}}}{1+\bar{\boldsymbol{\gamma}}_n}\ln\bar{\boldsymbol{\gamma}}_n$.
10: Update $n=n+1$.
11: **until** $n=N$ or convergence.
12: Output: The optimal solution to the problem (30), i.e., $\boldsymbol{\beta}^*$ and R_{sum}^*.

VI. Simulation Results and Discussion

In this section, simulation results are provided to verify the performance of the proposed algorithms. Set $K_{cu}=100$, $K_{su}=20$, $L=10$, $N_t=16$, $B=5$bits and $\sigma^2=-110$ dBm, unless otherwise stated. The number of propagation paths $N=3$ and the path-loss coefficient at the reference distance $\rho_0=20\lg(\frac{4\pi f_c}{c})$ dB with $f_c=28$ GHz and $c=3\times10^8$ m/s. $\alpha_{LOS}=2$ and $\alpha_{NLOS}=3$. For simplicity, we assume that $|Q_m|=Q$, $r_j^{[m]}=r=0.1$ bit/s/Hz and $r_{su}^{[m]}=r_{su}=8$ bit/s/Hz for $m\in\mathcal{M}$. The radius of the area of the legitimate users, SUs and eavesdroppers are set as $R_D=20$ m, $R_{su}=10$

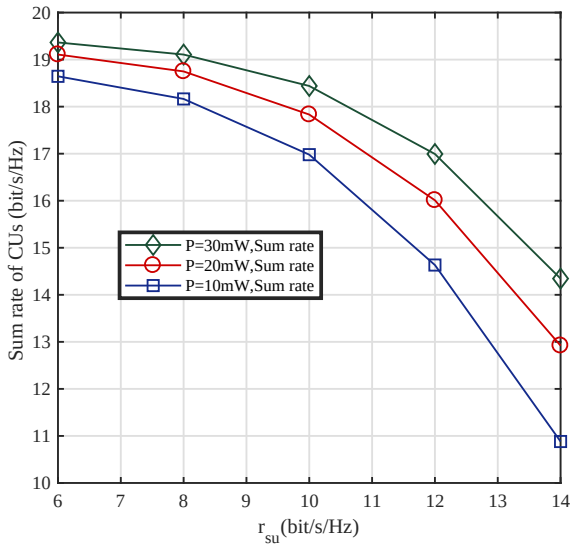


Fig. 2. Performance of the sum rate of CUs under different data rate thresholds of SUs. $P = 10$ mW, $P = 20$ mW and $P = 30$ mW, respectively.

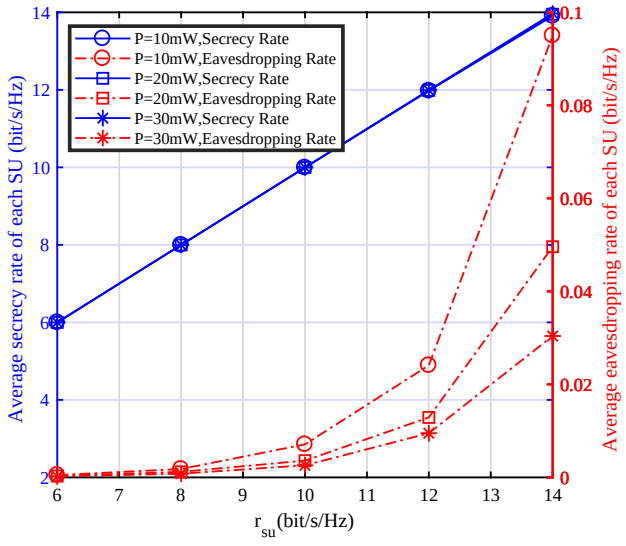


Fig. 3. Performance of the average secrecy rate and eavesdropping rate of each SU under different data rate thresholds of SUs. $P = 10$ mW, $P = 20$ mW and $P = 30$ mW, respectively.

m, and $R_E = 15$ m, respectively, and the legitimate users, SUs and eavesdroppers are uniformly distributed. In addition, we define the feasibility probability as the ratio of the number of channel realizations that the problem (30) can be solved to the total number of channel realizations, and the problem is always feasible unless otherwise stated.

First, the impact of the data rate requirements of SUs on the transmission performance of CUs, SUs and eavesdroppers are investigated in Figs. 2-3. From Fig. 2, it can be seen that the sum rate of CUs decreases as the rate threshold increases, while it will increase with the growth of transmit power. This is because more power will be assigned to the SUs to satisfy their increasing rate demands. In Fig. 3, it is easy to find that the interception performance of the eavesdroppers is severely deteriorated with the intra-group and inter-group interference, and the eavesdropping rate will further decline when the transmit power increases. Therefore, the average secrecy rate of each SU is rising with the increasing r_{su} and nearly close to the data rate thresholds.

Figs. 4 and 5 show how the sum rate of CUs and the average secrecy rate of each SU are influenced by the transmit power, the number of users in each group and the number of antennas at the BS. In Fig. 4, the sum rate of CUs becomes higher with the increasing transmit power, which is the same as the results in Fig. 2. In addition, as the number of users assigned to each group increases, the growth of the sum rate of CUs is not noticeable due to the existence of interference. However, when the number of transmit antennas increases, the sum rate will be significantly raised. This reveals that more users can be adopted in each group when antennas are abundant, yet the complexity of SIC will increase. In addition, as shown in Fig. 5, the average secrecy rate of each SU becomes higher with the transmit power due to the intra-group and inter-group interference, and the increasing number of users in each group has little impact on the security. As expected, the information leakage will further decrease with more antennas equipped at the BS.

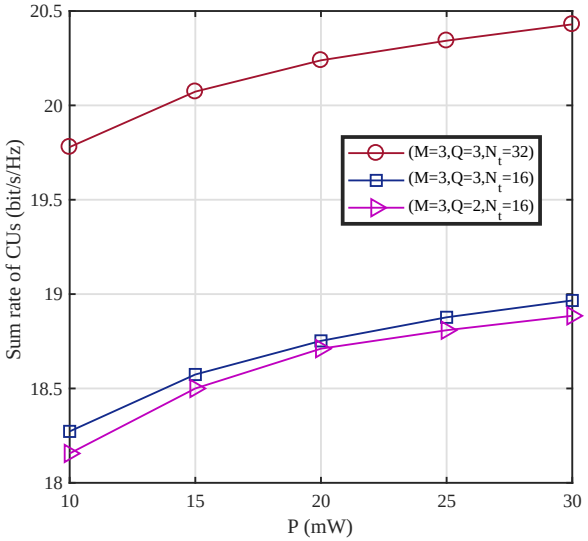


Fig. 4. Performance of the sum rate of CUs under different power consumption thresholds. Three cases are considered, i.e., $Q = 3$, $N_t = 32$, $Q = 3$, $N_t = 16$ and $Q = 2$, $N_t = 16$.

In Fig. 6, the influence of the number of groups and eavesdroppers on the average secrecy rate of each SU is studied. $K_{su} = 50$. From the results, we can see that the average secrecy rate is reduced with the number of groups due to the fact that the probability of confidential information leakage becomes larger when the number of SUs increases. Moreover, the growth of the number of eavesdroppers is also harmful to the security, and the downtrend becomes more obvious with more SUs. Nonetheless, the results also show that the amount of the information intercepted by the eavesdroppers is really small with the confusion caused by the intra-group and inter-group interference in dense networks.

In the Figs. 7 and 8, we discuss the impact of different grouping methods on the sum rate of CUs and average secrecy rate of each SU. Herein, “As per channel correlation”

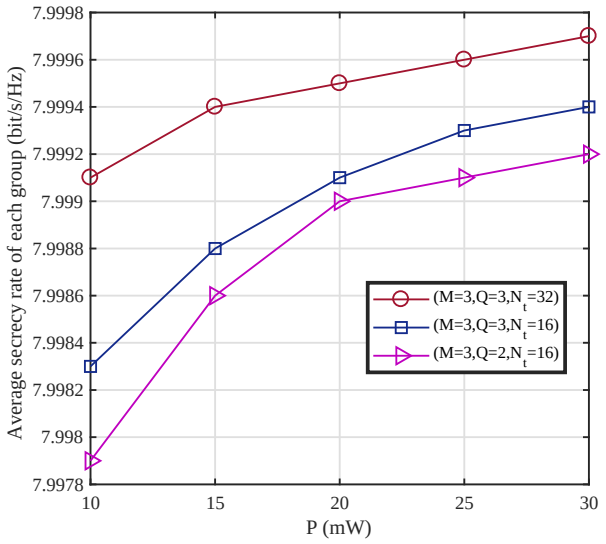


Fig. 5. Performance of the average secrecy rate of each SU under different power consumption thresholds. Three cases are considered, i.e., $Q = 3$, $N_t = 32$, $Q = 3$, $N_t = 16$ and $Q = 2$, $N_t = 16$.

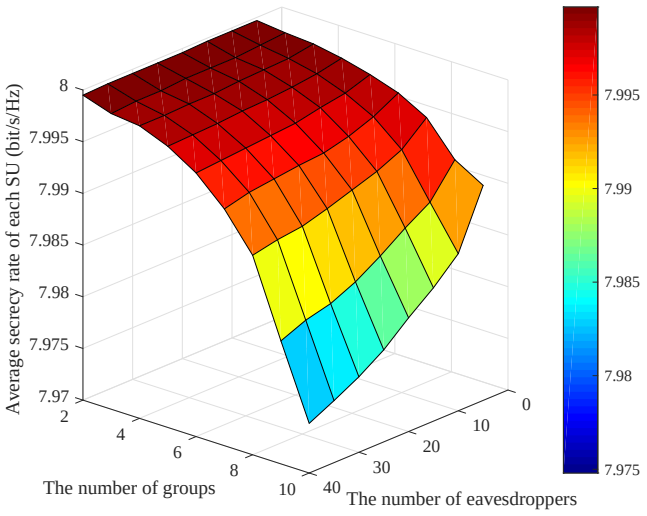


Fig. 6. Performance of the average secrecy rate of each SU with the increasing number of groups and eavesdroppers. Q=2.

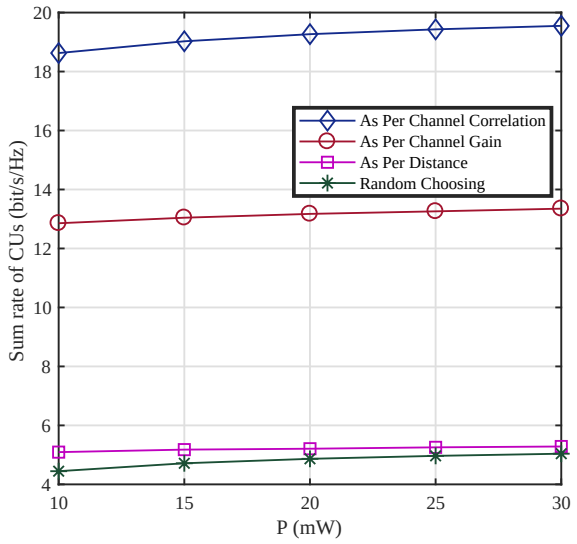


Fig. 7. Performance of sum rate of CUs under varying power consumption thresholds with different grouping methods. M=3, Q=2.

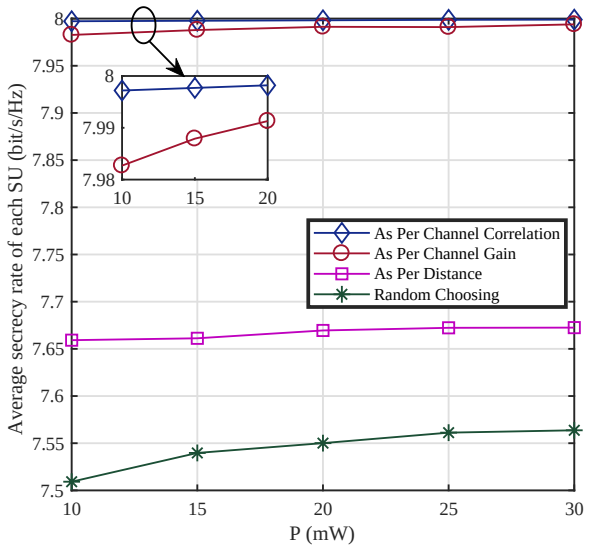


Fig. 8. Performance of average secrecy rate of each SU under varying power consumption thresholds with different grouping methods. M=3, Q=2.

corresponds to the method adopted in this paper. "As per channel gain" means that the users with stronger channel gains are selected as SUs and CUs. "As per distance" means that the users closer to the BS are chosen as SUs and CUs, and the nearer SU is paired with the nearer CU. "Random choosing" denotes that we randomly choose the SUs and CUs among all the users. In this simulation, the feasibility of the power optimization is considered due to the fact that the problem is not always solvable in the benchmark grouping methods. In Fig. 7, it can be observed that the proposed grouping algorithm has the optimal performance of sum rate of CUs compared to the other benchmarks. This is because the inter-group interference is effectively mitigated and the received signal power is significantly improved using the channel correlation. As for the secrecy capacity, the performance of the proposed grouping algorithm is still superior to all the other grouping schemes. Furthermore, it is proved that the advantage of distance is weaken in the multiple-antenna systems. In conclusion, the proposed grouping algorithm in this work performs well in terms of the CUs' sum rate and SUs' secrecy rate.

Finally, two benchmark schemes are provided to validate the effectiveness of the proposed scheme. In the mmWave-OMA scheme, all the users in each group are accessed via TDMA, and the similar optimization problem is formulated. For the scheme "Not-strongest-SU", we consider the second strongest user as the secure one without changing the hybrid beamforming, i.e., the SUs in this case also suffer from intra-group and inter-group interference. In addition, the terrible case that there exists at least one eavesdropper aligning to the direction of SUs is considered to examine the secrecy capacity, i.e., at last one eavesdropping channel are highly correlated with the confidential channel of SU in each group. Likewise, the feasibility of the optimization problem is also

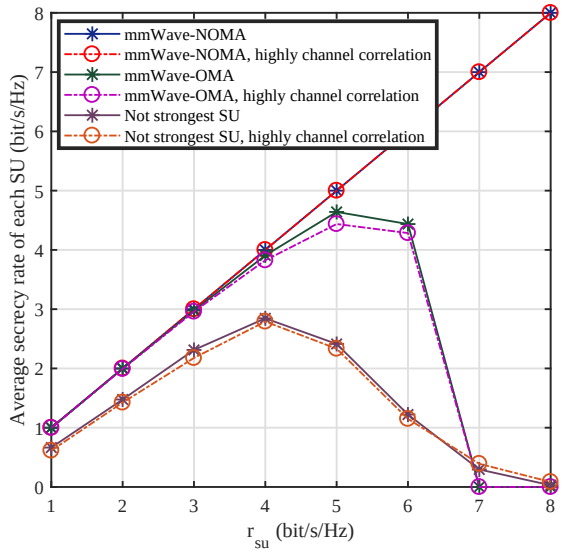


Fig. 9. Performance comparison of average secrecy rate of each SU in different schemes under varying SUs' data rate thresholds. The highly channel correlation between the SUs and eavesdroppers are considered. M=3, Q=3.

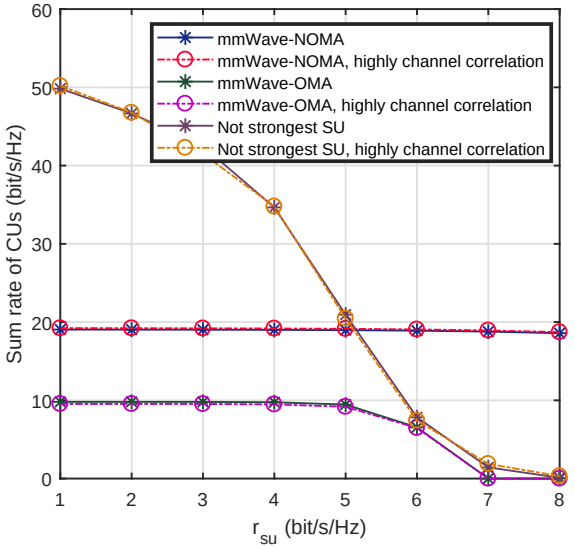


Fig. 10. Performance comparison of sum rate of CUs in different schemes under varying SUs' data rate thresholds. The highly channel correlation between the SUs and eavesdroppers are considered. M=3, Q=3.

considered. From Fig. 9, it is seen that the secrecy performance of proposed scheme in this work is superior to the benchmark schemes, and the highly channel correlation nearly has no impact on the secrecy rate thanks to the existence of intra-group and inter-group interference at the eavesdroppers. However, the mmWave-OMA and "Not-strongest-SU" schemes cannot achieve higher secrecy rate due to the lower spectrum efficiency and weakened legitimate channels quality, respectively. Their secrecy capacity is also apparently degraded in the case of highly channel correlation. In Fig. 10, actually, the highly channel correlation has no impact on the performance of CUs' sum rate since the eavesdropping CSI is not considered in the optimization problem. Furthermore, the results show that the sum rate of CUs in the mmWave-NOMA scheme is higher than that in the mmWave-OMA scheme due to high spectrum efficiency of NOMA. It is worth pointing out that the sum rate of CUs obtained in the "Not-strongest-SU" scheme is the highest at lower rate thresholds of SUs yet severely reduced when r_{su} becomes larger, which implies that the proposed mmWave-NOMA scheme makes a compromise between the CUs' sum rate and SUs' secrecy rate.

VII. Conclusions

In this paper, we present a novel framework for the mmWave-NOMA system to provide diverse services, where the users with higher channel strength are regarded as SUs, and the remaining ones are served as CUs. The hybrid precoding is specially designed to enhance the secrecy performance of SUs. To reduce the complexity and guarantee the quality of heterogeneous services, user grouping and PA are jointly optimized to maximize the sum rate of CUs while satisfying the rate threshold of the SUs. The non-convex problem is decomposed into two sub-problems, i.e., a hybrid SU-CU grouping algorithm based on matching theory is exploited to realize the optimal user grouping, and a low-complexity SCA-based algorithm is proposed to achieve power optimization whereby Lagrangian dual decomposition. Simulation results validate that both the security of SUs and the throughput of CUs can be improved using the proposed algorithms.

References

[1] Y. Cao, S. Wang, M. Jin, N. Zhao, Y. Chen, Z. Ding, and X. Wang, "Power optimization for secure mmWave-NOMA network with hybrid SU-CU grouping," in *Proc. IEEE GLOBECOM'21*, pp. 1–6, Madrid, Spain, Dec. 2021.

[2] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quart., 2018.

[3] M. Vaezi, R. Schober, Z. Ding, and H. V. Poor, "Non-orthogonal multiple access: Common myths and critical questions," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 174–180, Oct. 2019.

[4] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[5] Y. Liu, H. Chen, and L. Wang, "Physical layer security for next generation wireless networks: Theories, technologies, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 347–376, 1st Quart. 2017.

[6] R. Li, L. Wang, X. Tao, M. Song, and Z. Han, "Generalized benders decomposition to secure energy-efficient resource allocation for multiuser full-duplex relay cooperative networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 10728–10741, Aug. 2019.

[7] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

[8] F. Zhou, Z. Chu, H. Sun, R. Q. Hu, and L. Hanzo, "Artificial noise aided secure cognitive beamforming for cooperative MISO-NOMA using SWIPT," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 4, pp. 918–931, Apr. 2018.

[9] L. Lv, Z. Ding, Q. Ni, and J. Chen, "Secure MISO-NOMA transmission with artificial noise," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6700–6705, Mar. 2018.

[10] Y. Feng, S. Yan, C. Liu, Z. Yang, and N. Yang, "Two-stage relay selection for enhancing physical layer security in non-orthogonal multiple access," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 6, pp. 1670–1683, Jun. 2019.

[11] Y. Cao, N. Zhao, G. Pan, Y. Chen, L. Fan, M. Jin, and M. Alouini, "Secrecy analysis for cooperative NOMA networks with multi-antenna full-duplex relay," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5574–5587, Aug. 2019.

[12] N. Zhao, D. Li, M. Liu, Y. Cao, Y. Chen, Z. Ding, and X. Wang, "Secure transmission via joint precoding optimization for downlink MISO NOMA," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7603–7615, Aug. 2019.

[13] Y. Cao, N. Zhao, Y. Chen, M. Jin, Z. Ding, Y. Li, and F. R. Yu, "Secure transmission via beamforming optimization for NOMA networks," *IEEE Wireless Commun.*, vol. 27, no. 1, pp. 193–199, Feb. 2020.

[14] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656–1672, Mar. 2017.

[15] Z. Ding, Z. Zhao, M. Peng, and H. V. Poor, "On the spectral efficiency and security enhancements of NOMA assisted multicast-unicast streaming," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 3151–3163, Jul. 2017.

[16] X. Chen, Z. Zhang, C. Zhong, D. W. K. Ng, and R. Jia, "Exploiting inter-user interference for secure massive non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 4, pp. 788–801, Apr. 2018.

[17] K. Xiao, L. Gong, and M. Kadoch, "Opportunistic multicast NOMA with security concerns in a 5G massive MIMO system," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 91–95, Mar. 2018.

[18] Y. Xu, T. Zhang, D. Yang, Y. Liu, and M. Tao, "Joint resource and trajectory optimization for security in UAV-assisted MEC systems," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 573–588, Jan. 2021.

[19] X. Pang, J. Tang, N. Zhao, X. Zhang, and Y. Qian, "Energy-efficient design for mmWave-enabled NOMA-UAV networks," *Sci. China Inf. Sci.*, vol. 64, no. 4, Art. no. 140303, Apr. 2021.

[20] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, Feb. 2017.

[21] J. Cui, Y. Liu, Z. Ding, P. Fan, and A. Nallanathan, "Optimal user scheduling and power allocation for millimeter wave NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1502–1517, Mar. 2018.

[22] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X. Xia, "Millimeter-wave NOMA with user grouping, power allocation and hybrid beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5065–5079, Nov. 2019.

[23] K. Wang, J. Cui, Z. Ding, and P. Fan, "Stackelberg game for user clustering and power allocation in millimeter wave-NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 2842–2857, May 2019.

[24] J. Ren, Z. Wang, M. Xu, F. Fang, and Z. Ding, "An EM-based user clustering method in non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8422–8434, Dec. 2019.

[25] F. Zhao, W. Hao, L. Shen, G. Sun, Y. Zhou, and Y. Wang, "Secure energy efficiency transmission for mmwave-NOMA system," *IEEE Syst. J.*, vol. 15, no. 2, pp. 2226–2229, Jun. 2021.

[26] S. Huang, M. Xiao, and H. V. Poor, "On the physical layer security of millimeter wave NOMA networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11697–11711, Oct. 2020.

[27] X. Sun, W. Yang, and Y. Cai, "Secure communication in NOMA-assisted millimeter-wave SWIPT UAV networks," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 1884–1897, Mar. 2020.

[28] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. on Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[29] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

[30] L. Dai, B. Wang, M. Peng, and S. Chen, "Hybrid precoding-based millimeter-wave massive MIMO-NOMA with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 131–141, Jan. 2019.

[31] E. Bodine-Baron, C. Lee, A. Chong, B. Hassibi, and A. Wierman, "Peer effects and stability in matching markets," *in Proc. Int. Symp. Algorithmic Game Theory*, pp. 117–129, Jul. 2011.

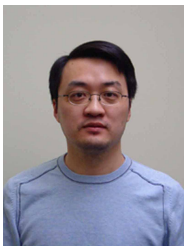
[32] Z. Han, Y. Gu, and W. Saad, *Matching Theory for Wireless Networks*. Berlin: Springer, 2017.

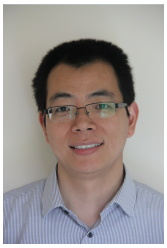
[33] J. Papandriopoulos and J. S. Evans, "SCALE: a low-complexity distributed protocol for spectrum balancing in multiuser DSL networks," *IEEE Trans. Inform. Theory*, vol. 55, no. 8, pp. 3711–3724, Aug. 2009.

[34] S.Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge: U.K.: Cambridge Univ. Press, 2004.

[35] X. Zhang, X. Tao, Y. Li, N. Ge, and J. Lu, "On relay selection and subcarrier assignment for multiuser cooperative OFDMA networks with QoS guarantees," *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4704–4717, Apr. 2014.

[36] D. T. Ngo, S. Khakurel, and T. Le-Ngoc, "Joint subchannel assignment and power allocation for OFDMA femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 342–355, Dec. 2014.





Committee.